Classification Trees: A Data Mining Tool for Improved Sensitivity Analyses of Large-Scale Groundwater Models

Srikanta Mishra and Neil E. Deeds
INTERA Inc.
9111A Research Blvd, Austin, TX 78758
Tel – (+1) 512-425-2071 ; E-mail – smishra@intera.com

Groundwater availability models (GAMs) are complex regional-scale, distributed-parameter and integrated surface/groundwater models. Such models are currently being developed by the Texas Water Development Board for the major aquifers of Texas to provide a tool to estimate groundwater availability for various water use strategies and to determine the cumulative effects of increased water use and drought. For these complex nonlinear models with a large number of inputs, establishing input-output relationships using standard statistical techniques such as multivariate linear regression is not always possible. In such situations, classification tree analysis has been found to be a useful "data mining" tool for providing insights into what variable or variables are most important in determining whether outputs fall into different categories.

A binary decision tree is at the heart of classification tree analysis. The decision tree is generated by recursively finding the variable splits that best separate the output into groups where a single category dominates. For each successive fork of the binary decision tree, the algorithm searches through the variables one by one to find the purest split within each variable. The splits are then compared among all the variables to find the best split for that fork. The process is repeated until all groups contain a single category. In general, the variables that are chosen by the algorithm for the first several splits are most important, with less important variables involved in the splitting near the terminal end of the tree.

The use of classification trees in sensitivity analysis involves several steps beyond the basic tree construction. After the tree is built, nodes are evaluated as to their relative contribution in determining important variables. The earliest splits contribute most to the reduction in deviance and are considered to be most important in the classification process. The later splits may be marginally important, or may simply fall in the range of statistical "noise." Usually, an attempt is made to "prune" the tree (i.e., reduce the number of splits) to the point where only a handful of variables are left which can be used to classify the majority of the outputs. Pruning is usually accomplished by increasing the minimum reduction in deviance necessary for node splitting and then rebuilding the tree.

Traditional applications of classification trees have primarily been in the fields of medical decision making and data mining for social sciences. In this work, we discuss the basics of classification tree analysis and present two different applications of the methodology to results from a regional GAM for the Carrizo-Wilcox aquifer in southwest Texas.

The Carrizo-Wilcox aquifer is comprised of hydraulically connected sands from the Wilcox Group and the Carrizo Formation of the Claiborne Group, and is classified as a major aquifer in Texas. The model area for the southern Carrizo-Wilcox GAM is bounded laterally on the northeast by the surface water basin divide between the Guadalupe and Colorado Rivers and to the southwest by the Rio Grande River. The model grid for the Southern Carrizo-Wilcox GAM is 217 rows by 112 columns with 6 layers. All GAMs are required to be developed using MODFLOW96, with a regular grid spacing of 1 mile square.

In the first application, we use classification tree analyses to determine the key drivers of extreme head residuals from a single deterministic simulation. This analysis is completed on head residuals after the model has been calibrated. **Figure 1a** shows a scatterplot of measured and simulated heads for the southern Carrizo-Wilcox GAM. A prerequisite to classification tree analysis is the division of the dependent variable into two classes. For this case, we divided the head residuals (the difference between the measured and model-predicted head) into "small" and "large" categories (**Figure 1b**). The "small" category consisted of residuals that fell below the 10% residual value. The "large" category consisted of residuals that were larger than the 90% residual value. The input variables considered in the analysis were stress period, well location easting and northing, measured value, simulated value, and drawdown from steady-state. With stress period, the intent was to capture the importance of time. The well location easting and northing provided a measure spatial variability. The measured and simulated values provided insight into calibration bias. Finally, the drawdown from steady-state was a variable that indicated the amount of stress historically existing in that region of the model.

The classification tree analysis classified 91.8% of the residuals correctly after only two binary splits. The analysis indicated that the most important variable was the drawdown from steady-state. The primary importance of this variable is indicated by it being chosen for the first split. The second most important variable was the well location easting. **Figure 2** shows a partition plot with these two variables. The analysis indicates that the largest errors occurred in the western area of the model where the most stress had historically occurred in the aquifer.

In the second application, we use classification tree analyses to help identify which variables or combinations of variables are most important in model calibration. Insights into the range of parameter uncertainties were obtained from trial-and-error calibrations supplemented by one-at-a-time perturbation-based sensitivity analyses. These uncertainty ranges were used as input to Monte Carlo simulation runs. Model outcomes were partitioned into those that matched the data at some level of acceptance (category=FIT) and those that do not (category=MISFIT). The classification tree algorithm was then applied to determine the key parameters responsible for driving model outcomes into the FIT or MISFIT categories. The most important parameters differed between the steady-state and transient models. In general, for the steady-state model, recharge and vertical hydraulic conductivity were most important. For the transient model, horizontal hydraulic conductivity and pumping were most important.

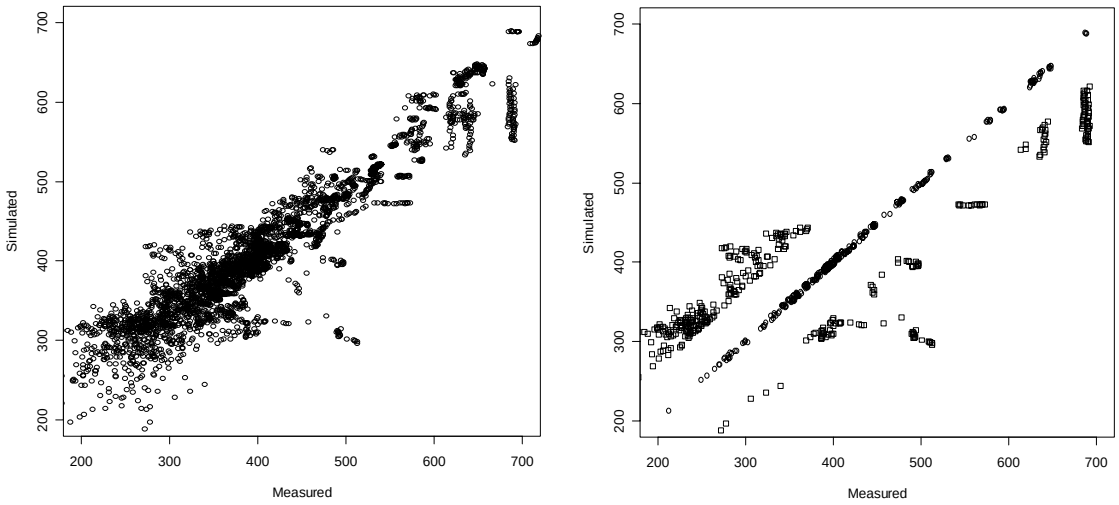

Figure 1: Scatterplots of (a) all measured versus simulated heads and (b) measured and simulated heads comprising the top and bottom 10% head residuals.

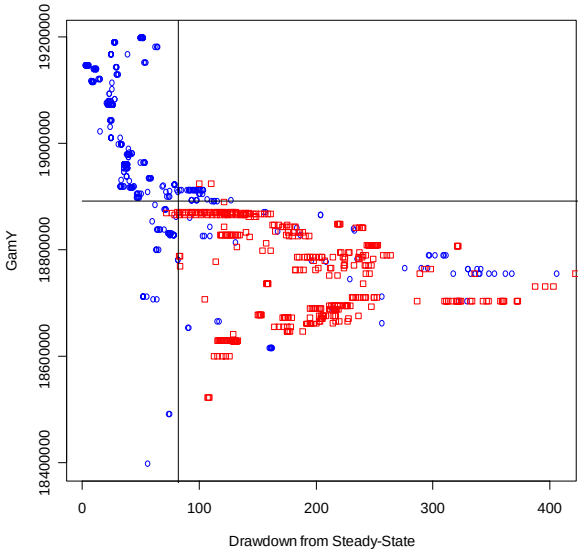

Figure 2: Partition plot of the top two variables: drawdown from steady-state and well location easting.